BIOINFORMATICS ORIGINAL PAPER
Vol. 21 no. 3 2005, pages 333–337
doi:10.1093/bioinformatics/bti008

Evaluating putative chimeric sequences from PCR-amplified products

Juan M. Gonzalez, Johannes Zimmermann and Cesareo Saiz-Jimenez*

Instituto de Recursos Naturales y Agrobiologia, CSIC, Apartado 1052, 41080 Sevilla, Spain

Received on August 4, 2004; revised and accepted on August 30, 2004
Advance Access publication September 3, 2004

ABSTRACT

Motivation: PCR amplification of highly homologous genes from complex DNA mixtures is known to generate a significant proportion of chimeric sequences. Ribosomal RNA genes are used for microbial species detection and identification in natural environments, and current assessments of microbial diversity are based on these sequences. Thus, chimeric sequences could lead to the discovery of non-existent microbial species and false diversity estimates.

Methods: In essence, our only source of information to decide if a sequence is chimeric or not is to compare it with known, non-chimeric sequences. Putative chimeric sequences were analyzed from sequence fragments of selected length (referred to as words) by comparing nucleotides at corresponding positions. Distances for each word between reference sequences (closely related to the tested sequence) were compared to the differences introduced by the tested sequence. The proposed strategy considers the actual variability existing in different regions throughout the analyzed sequences. The result is an efficient strategy for the evaluation of putative chimeric sequences.

Availability: A program computing the above procedure, Chimera and Cross-Over Detection and Evaluation (Ccode), is available at http://www.irnase.csic.es/users/jmgrau/index.html and http://www.rtphc.csic.es/download.html

Contact: jmgrau@irnase.csic.es

INTRODUCTION

Advances in environmental microbiology have generated a completely new perspective on microbial diversity (Ward *et al*., 1990; Curtis *et al*., 2002; DeLong, 2001). In fact, an astonishing number of novel candidate bacterial divisions are being proposed based solely on PCR-amplified 16S rRNA gene sequences retrieved from environmental samples (Hugenholtz *et al*., 1998; Pace, 1997). PCR amplification is the standard means of detecting and identifying microorganisms in complex, natural environments. Amplification biases and chimeric sequences have been reported to occur during DNA amplification by PCR from mixtures of sequences, such as environmental DNA samples (von Wintzingerode, 1997; Suzuki and Giovannoni, 1996). Chimeras are usually PCR artifacts resulting from a prematurely terminated amplicon when it reanneals to a different template DNA and is copied to completion based on this second parental sequence (Wang and Wang, 1996). A chimeric sequence, or chimera, is composed of two or more phylogenetically distinct parental sequences. Chimeras are a serious concern in culture-independent surveys of microbial communities because they suggest the presence of non-existing microorganisms (von Wintzingerode *et al*., 1997), above all if one considers that most microorganisms in nature are unculturable (Ward *et al*., 1990; Pace, 1997). The occurrence of chimeric sequences weakens the base of the currently accepted model and of the evidence it has produced for a large microbial diversity on our planet (Curtis *et al*., 2002; Hugenholtz *et al*., 1998; Ward *et al*., 1990).

In view of the above scenario, there is a need for computing initiatives capable of evaluating whether an amplified PCR product represents a chimera. Several methods have been proposed to detect chimeric sequences, such as different variants of the nearest-neighbor method (Robinson-Cox *et al*., 1995; Komatsoulis and Waterman, 1997; Cole *et al*., 2003), of which the most frequently used is 'Chimera Check' (Cole *et al*., 2003), or the recently introduced 'Bellerophon' (Huber *et al*., 2004). Most of these methods are based on the principle that a chimera would show different phylogenetic relationships depending on the part—beginning or ending—of the sequence to be analyzed. This approach has been successful in the detection of numerous chimeras both from natural studies (Robinson-Cox *et al*., 1995; Komatsoulis and Waterman, 1997) and from DNA databases (Hugenholtz and Huber, 2003). However, there is no strategy to decide whether those sequences are in fact chimerical or not. This study analyzes the problems involved in detecting and evaluating the chimeric sequences, suggesting an alternative approach based on known variabilities among related sequences.

*To whom correspondence should be addressed.

METHODS

Classifying a query sequence as chimeric or non-chimeric is not a simple matter. In essence, the problem can be reduced to the need to evaluate the added variability introduced using a query sequence within a set of reference sequences (the closest relatives to the query sequence). To analyze a putative chimeric sequence, a set of the closest sequences available in the databases should be obtained. A comparison of these reference sequences provides the variability within references, which is to be compared with the existing variability between query and reference sequences. These comparisons are performed on fragments of the full-length sequences and evaluation of the possible origin of these fragments should confirm or discount a chimeric origin for the sections of a full sequence. For any comparison among sequences, a reliable alignment is an absolute requirement.

A chimeric sequence is composed of at least two partial sequences from different real genes. Chimeric sequences comprising more than two partial sequences are frequently found, resulting in cross-over artifacts. In order to detect a differential origin between portions of sequences, sequences are examined by fragments (words). These fragments may be of a selected size (word length) depending on a number of factors such as type and length of the sequence to be analyzed. This approach is based on the differential variability between areas of aligned sequence sets, and so the results are independent of the existence of conserved or variable regions. Pairwise comparisons of aligned sequences are performed and the total distance per word is estimated for each. For a pairwise comparison, the distance value for a word (d) was obtained as a sum of differences:

$$d = \sum_{i}^{w} d_i ,$$

where w is the number of nucleotides composing a word or word length, and d_i denotes the number of differences at a given nucleotide position in a word for the aligned pairwise comparison.

Average distances for each word (avgR) are computed including every combination of pairwise comparisons among the selected reference sequences (n).

$$\text{avgR} = \sum_{j}^{n} \sum_{i}^{w} d_{ij} .$$

The same calculations are computed for distances among query and reference sequences (avgQ). Values of avgR and avgQ are obtained for each word forming the sequences under analysis.

Distances among reference sequences are expected to be lower than distances between query and reference sequences for each word belonging to a chimeric fragment. Similar distances should exist when the non-chimeric portion of a sequence is to be compared. Thus, the ratio avgQ/avgR should be equal to or greater than one (avgQ/avgR $\geq$ 1).

A decision on the chimeric/non-chimeric origin of a query sequence is adopted based on a 95% confidence limit of avgQ and a test of analysis of variance (Sokal and Rohlf, 1981). Each word is suggested to be a chimeric sequence fragment if avgR is lower than the confidence limit around avgQ. Confidence limits were calculated as avgQ $\pm$ t sdQ (Sokal and Rohlf, 1981), where avgQ and sdQ are the average and SD, respectively, and t is the t-Student critical value for $n - 1$ degrees of freedom where n is the total number of pairwise comparisons between query and reference sequences. A second criterium for suggesting that a word could have a chimeric origin is based on a test of analysis of variance (ANOVA). ANOVA is performed among two sets of data. One set represents distances between reference sequences and another is constituted by distances between query and reference sequences (Fig. 1).

A program written in C, Ccode (Chimera and Cross-Over Detection and Evaluation), performs the above procedure. Ccode is freely available at http://www.irnase.csic.es/users/jmgrau/index.html and http://www.rtphc.csic.es/download.html. The closest relatives to the query sequence were considered as reference sequences. Reference sequences were obtained using the blastn algorithm (Altschul *et al*., 1990) at the NCBI (http://www.ncbi.nlm.nih.gov/BLAST/). Multiple alignments were performed by clustalW1.82 (Thompson *et al*., 1994) followed by manual inspection of its results. Scripts are available at the URL address given above to automatize the process of alignment and chimera evaluation for multiple query sequences.

RESULTS

The protocol outlined in this report has been tested on a number of sequences. For example, the absence of chimeras among eighteen 16S rDNA sequences was confirmed during microbial surveys of Acidobacteria in hypogean environments (J.Zimmermann, J.M.Gonzalez, W.Ludwig and C.Saiz-Jimenez, submitted for publication; Table 1). Evaluation of the results provided by Chimera Check (Cole *et al*., 2003) on these sequences also suggested the absence of chimeras in that dataset. These sequences were confirmed to be non-chimeric after comparison with recently found sequences in other environments. In addition, we have performed a screening of the sequences from databases suggested as putative chimeras by Hugenholtz and Huber (2003). Among the 39 sequences suggested by these authors, we could confirm all of them as chimeric DNA using the procedure proposed in this study. Using the program Chimera Check (Cole *et al*., 2003), we could only detect 46% of the sequences proposed by Hugenholtz and Huber (2003) as chimeras. Thus, the proposed procedure (Ccode) has been successful in confirming a number of chimeras. As an example, difficulties were encountered in showing the chimeric origin of sequence

A.
Query sequence (Number of bases): AF068806 (1419)
Reference sequences (Number of bases; Percent similarity with Query):
AY225613 (1419; 90.3)
AY225615 (1419; 89.8)
AF367490 (1419; 89.4)
AB088431 (1419; 88.8)
AB088432 (1419; 88.7)
AF355050 (1419; 87.7)
U15104 (1419; 88.7)
U15100 (1419; 88.4)
Word Length = 141.

Bases	AvgQ (sdQ)	AvgR (sdR)	Ratio	Anova[1, 34]
141	10.62 (6.23)	10.29 (5.37)	1.03	0.023
282	8.12 (3.87)	9.86 (4.47)	0.82	0.985
423	0.25 (0.71)	0.50 (0.88)	0.50	0.540
564	1.75 (1.04)	2.43 (1.40)	0.72	1.614
705	3.62 (2.20)	4.43 (2.59)	0.82	0.636
846	6.88 (3.76)	5.18 (3.61)	1.33	1.350
987	8.38 (6.48)	8.89 (6.15)	0.94	0.043
1128	24.25 (0.46)	1.68 (2.18)	14.45*	831.600*
1269	41.38 (0.52)	0.75 (0.65)	55.17*	26601.852*
1410	36.38 (3.20)	9.04 (8.74)	4.03*	74.069*

Q -> Comparisons between query sequence and reference sequences.
R -> Comparisons between reference sequences.
Significance at P<0.05 level is indicated by *.
Number of pairwise comparisons between reference sequences = 28,
and between query and reference sequences = 8.
Results of ANOVA are given for the degrees of freedom between brackets.

B.
Query sequence (Number of bases): NC_000961 (1364)
Reference sequences (Number of bases; Percent similarity with Query):
Z54172 (1364; 99.6)
U20163 (1364; 98.5)
AY519654 (1364; 99.0)
AY099168 (1364; 99.2)
AJ419868 (1364; 99.6)
Word Length = 136.

Bases	AvgQ (sdQ)	AvgR (sdR)	Ratio	Anova[1, 4]
136	1.00 (1.73)	1.90 (1.85)	0.53	0.818
272	0.40 (0.89)	0.80 (1.03)	0.50	0.542
408	0.00 (0.00)	0.00 (0.00)	1.00	0.000
544	0.40 (0.89)	0.80 (1.03)	0.50	0.542
680	0.40 (0.89)	0.80 (1.03)	0.50	0.542
816	0.20 (0.45)	0.40 (0.52)	0.50	0.542
952	0.00 (0.00)	0.00 (0.00)	1.00	0.000
1088	0.20 (0.45)	0.40 (0.52)	0.50	0.542
1224	3.60 (1.52)	2.40 (2.50)	1.50	0.951
1360	4.80 (3.90)	5.40 (3.41)	0.89	0.094

Q -> Comparisons between query sequence and reference sequences.
R -> Comparisons between reference sequences.
Significance at P<0.05 level is indicated by *.
Number of pairwise comparisons between reference sequences = 10,
and between query and reference sequences = 5.
Results of ANOVA are given for the degrees of freedom between brackets.

Fig. 1. Representative examples of the output generated by Ccode for the evaluation of chimeric sequences. (**A**) The results for a chimeric sequence (AF068806) are presented; and in (**B**) a non-chimeric, real sequence (NC_000961, 16S rRNA gene) is examined. The output indicates the query and reference sequences used during the analysis. Between parentheses, following the sequence accession IDs, the number of aligned nucleotides considered in the analysis and the percentage of similarity to the query sequence are given. Word length can be selected as a percentage of full sequence (10% in the given examples) or as number of nucleotides. The results are provided for each word along the full aligned sequence. Estimated values for avgQ (average distance between query and reference sequences per word), avgR (average distance between reference sequences per word), their SD (in parentheses) and ratio (avgQ/avgR) are represented in columns for each word. An asterisk after the ratio value indicates whether a confidence limit for avgQ is significantly above the avgR values suggesting a sequence fragment with distinct origin. An ANOVA is computed and the results for each word are shown in the last column; an asterisk also indicates the significance above the $P < 0.05$ level for each word, suggesting the existence of a differential origin for that word. The example presented in (**A**) shows the results for a chimeric sequence composed of two fragments of different origin (0–987, and 988 to the end of the sequence).

AF253224 because the database contained a highly related and unreported chimeric sequence, AF253225, which had to be removed from the set of reference sequences previous to any screening for a chimera with the query sequence. A summary of the results on chimeric sequence detection using different strategies (Ccode, Chimera Ceck and Bellerophon) is reported in Table 1. In addition, a number of potential chimeras was indicated by using Bellerophon (Huber and Hugenholtz, 2004) and screened using Ccode and Chimera Check. Ccode (this study) was able to confirm ∼35% of sequences as chimeras while Chimera Check (Cole *et al*., 2003) only detected chimeras for ∼19% of the tested sequences. This confirms the complementarity, and non-exclusiveness, of the different chimera detection strategies.

DISCUSSION

In this study, we propose a strategy for evaluating chimeric sequences; it is based on the distances shown by fragments of a query sequence when compared to closely related reference sequences from databases, in the framework of pairwise comparisons among those reference sequences. It is assumed that selected reference strains limit the extent of variability allowed within a phyletic group. This variability is analyzed by words of a freely selectable length, so foreign fragments can be detected. The detection is based on the added variability introduced by a query sequence; if the query sequence is a chimera, it would introduce high variability while a related reference sequence will only represent a minor added variation to the analysis. Both chimeras and erroneous PCR amplifications can be detected using this strategy, always with reference to the distance detected among the closest relatives from public databases. This procedure considers the variability specific to certain regions of the tested sequence type (i.e. rRNA gene sequences) since both conserved and variable regions are found in almost every known gene or DNA fragment and this is also the case for the rRNA genes (de Rijk *et al*., 1995).

A correct evaluation of chimeric sequences is influenced by the selection of adequate reference sequences and an

Table 1. Comparative results showing the percentage of detected chimeras for various sets of sequences

Sequence dataset	Chimera detecting program Ccode	Chimera Check[a]	Bellerophon[b]
Hugenholtz and Huber (2003)[c]	100	46	100
Authors' unpublished data[d]	0	0	—
Database putative chimeras[e]	35	19	—

Chimera detection was performed by using three independent methods: Chimera check (Cole *et al*., 2003), Bellerophon (Huber and Hugenholtz, 2004) and Ccode (this study).
[a]After manual inspection and evaluation of results.
[b]Bellerophon requires a set of sequences belonging to the same DNA library to search for potential chimeras.
[c]Chimeric sequences proposed by Hugenholtz and Huber (2003) using Bellerophon. A total of 39 16S rDNA sequences.
[d]Sequences considered as non-chimeras from Authors' unpublished data. A total of 18 16S rDNA sequences. Accession numbers from AY703458 to AY703475.
[e]Putative chimeric sequences related to those proposed by Huber and Hugenholtz (2004) and proposed as chimeras by bellerophon. A total of 37 16S rDNA sequences. These sequences had the following accession numbers: AJ515717^{+}, AY234728^{++}, AF498724^{++}, AF498753^{++}, AJ581627, AJ347029^{++}, AJ347052^{++}, AJ347049^{++}, Y12597^{+}, Z95719^{+}, AJ347774, AB050205, AJ535118*, AB050229^{+}, AB050207^{+}, AJ309654^{+}, AY225613, AB058907, AB058908, AB058909, AB058910, AB058911, AB058914, AB058915, AB058916, AB058917, AB058918, AB100005^{+}, AY171615, AY337603^{+}, AF293010^{+}, AF293013^{+}, AJ224042, AJ224039, AF510191, AF353208, and AF422677^{+}. Those labeled with a simple and a double cross were suggested as chimeric by Ccode and Chimera Check, respectively. Asterisk indicates both methods detected a chimera.

accurate multiple sequence alignment. Reference sequences should represent the closest relatives to the query sequence indicating the acceptable range of variability in the phylogenetic group to be considered. It is advisable to ensure the absence of chimeric sequences within the reference sequence set since they would invalidate the analysis by introducing extra variability notwithstanding the real distances existing within the phylogenetic group being considered. The existence of chimeric sequences in public DNA databases is known (Hugenholtz and Huber, 2003), although the development of novel strategies for the detection and evaluation of chimeric sequences (Huber *et al*., 2004 and this study) will hopefully overcome this drawback. As with any comparative analysis to be performed among sequences, an alignment ensuring accurate base-to-base comparisons is of outmost importance. The results generated from poorly aligned sequences will lack any significance. Thus, we recommend manual inspection and editing of the alignments before any decision on the chimeric nature of a sequence is reached.

The program performing the strategy for chimera evaluation proposed in this study can analyze sequences for any required word length. Generally, values of 5–20% of sequence length appear to deliver accurate results, for example, working on 16S rDNA sequences with a full-length of ~1500 nt. It should be noted that the use of fragments either too long or too short might result in a reduction of sensitivity.

Several strategies for the detection of chimeric sequences have been proposed (Robinson-Cox *et al*., 1995; Komatsoulis and Waterman, 1997; Cole *et al*., 2003; Huber *et al*., 2004). They are based on the nearest-neighbor method that detects a chimera by comparative phylogenetic results obtained from two sequence fragments belonging to the initial and final portions of the tested sequence (Robinson-Cox *et al*., 1995). Currently, the most frequently used software is 'Chimera Check' (Cole *et al*., 2003). Recently, a new approach has been proposed, 'Bellerophon' (Huber *et al*., 2004), which is useful for analyzing the sequences obtained from single DNA libraries. The strategy presented in this study complements previous methods for chimera detection since it allows evaluation of the chimeric nature of a tested sequence. It performs an in-depth analysis on putative chimeric sequences and considers their closest relatives as well as the variability within their phylogenetic surroundings to classify a sequence as chimerical or not. Previous strategies for chimera detection (i.e. Chimera Check and Bellerophon) (Cole *et al*., 2003; Huber *et al*., 2004) provide results that require further evaluation by the researcher. In this study, the proposed strategy, performed by Ccode, provides tests of significance leading to a simple discrimination of chimera sequences.

The existence of a too diverse reference set of sequences is likely to impact negatively on meaningful detection of chimeric sequences by any proposed computational method. Closely related sequences, which could be adequate candidates for reference sequences, often show relatively high percentages of similarity over their full sequence length [as provided by the Blast algorithm (Altschul *et al*., 1990)]. Chimeric sequences frequently exhibit percentages of similarity (over full sequence length) to closest relatives around the species threshold [97%; Roselló-Mora and Amann (2001)]. Thus, considering as putative chimeras only those sequences showing similarity percentages below 97% (e.g. Chelius and Moore, 2004) is a precarious assumption.

Although sequence variability within phylogenetic groups is the only existing reference for assessing whether or not a sequence has a chimeric origin or is the result of crossing-over having occurred during PCR, the use of the known biodiversity as a tool for further analysis might introduce potential analytical problems. At present, a large portion of the biodiversity on our planet is known but it has been suggested that organisms yet to be discovered represent a major fraction of total microbial richness (Curtis *et al*., 2002). Thus, the existence of unknown diversity could imply a reduced

set of the actual variability for evaluating a chimera; this could lead to the classification of a sequence as a chimera that might simply be among the unknown, but actual, biodiversity. This selection of false positives appears as a minor error in today's growing DNA databases, but it needs to be considered, since the selection of non-chimeras as chimeric sequences could impede progress in understanding the actual diversity existing on the planet. Nevertheless, environmental molecular surveys are rapidly expanding DNA databases (i.e. Cole *et al*., 2003) and the possible problem will be significantly diminished over time.

Besides the potential challenges reported above, at present, there is a clear need for chimera-evaluating initiatives (von Wintzingerode *et al*., 1997; Cole *et al*., 2003; Hugenholtz and Huber, 2003 and this study). The risk involved in accepting chimeric sequences representing non-existing organisms is far higher than the possibility of discriminating some non-chimeric sequences in the process. DNA amplification by PCR is the basis for the analyses performed during environmental molecular biodiversity surveys (Ward *et al*., 1990; Pace, 1997; von Wintzingerode *et al*., 1997), and so the risks due to PCR-derived artifacts are continuously increasing. Thus, the present and future initiatives to detect and evaluate putative chimeric sequences are required and should complement any molecular biodiversity survey to be carried out on environmental samples.

CONCLUSION

This study reports a novel strategy and computer program for the evaluation of chimeric sequences that complements previous software and methodologies. The method overcomes the need for manual inspection of putative chimeric sequences and avoids the application of a subjective or biased personal perspective to the evaluation of putative chimeric sequences. A program performing the proposed strategy is available on the Web.

ACKNOWLEDGEMENTS

The authors thank the helpful assistance of Dr Matthias Keil in porting Ccode to the Windows platform and Dr Adrian Pearce for his helpful comments on the manuscript. The authors acknowledge support through projects REN2002-00041, REN2003-02854 and BTE2002-04492-C02-01 from the Spanish Ministry of Education and Science (MEC). J.M.G. and J.Z. were supported by the MEC (Ramon y Cajal Programme) and the Marie Curie Programme, respectively.

REFERENCES

Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.

Chelius,M.K. and Moore,J.C. (2003) Molecular phylogenetic analysis of Archaea and Bacteria in Wind Cave, South Dakota. *Geomicrobiol. J*., **21**, 123–134.

Cole,J.R., Chai,B., Marsh,T.L., Farris,R.J., Wang,Q., Kulam,S.A., Chandra,S., McGarrell,D.M., Schmidt,T.M., Garrity,G.M. and Tiedje,J.M. (2003) The Ribosomal Database Project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy. *Nucleic Acids Res*., **31**, 442–443.

Curtis,T.P., Sloan,W.T. and Scannell,J.W. (2002) Estimating prokaryotic diversity and its limits. *Proc. Natl Acad. Sci. USA*, **99**, 10494–10499.

DeLong,E.F. (2001) Microbial seascapes revisited. *Curr. Opin. Microbiol*., **4**, 290–295.

de Rijk,P., Van de Peer,Y., Van den Broeck,I. and de Wachter,R. (1995) Evolution according to large ribosomal subunit RNA. *J. Mol. Evol*., **41**, 366–375.

Huber,T., Faulkner,G. and Hugenholtz,P. (2004) Bellerophon: a program to detect chimeric sequences in multiple sequence alignments. *Bioinformatics*, DOI: 10.1093/bioinformatics/bth226.

Hugenholtz,P. and Huber,T. (2003) Chimeric 16S rDNA sequences of diverse origin are accumulating in the public databases. *Intl. J. Syst. Evol Microbiol*., **53**, 289–293.

Hugenholtz,P., Goebel,B.M. and Pace,N.R. (1998) Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *J. Bacteriol*., **180**, 4765–4774.

Komatsoulis,G.A. and Waterman,M.S. (1997) A new computational method for the detection of chimeric 16S rRNA mixed bacterial populations. *Appl. Environ. Microbiol*., **63**, 2338–2346.

Pace,N.R. (1997) A molecular view of microbial diversity and the biosphere. *Science*, **276**, 734–740.

Robinson-Cox,J.F., Bateson,M.M. and Ward,D.M. (1995) Evaluation of nearest-neighbor methods for the detection of chimeric small-subunit rRNA sequences. *Appl. Environ. Microbiol*., **61**, 1240–1245.

Roselló-Mora,R. and Amann,R. (2001) The species concept for prokaryotes. *FEMS Microbiol. Rev*., **25**, 36–67.

Sokal,R.R. and Rohlf,F.J. (1981) *Biometry*, 2nd edn. W.H. Freeman and Co., NY.

Suzuki,M.T. and Giovannoni,S.J. (1996) Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Appl. Environ. Microbiol*., **62**, 625–630.

Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix-choice. *Nucleic Acids Res*., **22**, 4673–4680.

von Wintzingerode,F., Göbel,U.B. and Stackebrandt,E. (1997) Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis. *FEMS Microbiol. Rev*., **21**, 213–229.

Ward,D.M., Weller,R. and Bateson,M.M. (1990) 16S rRNA sequences reveal numerous uncultured microorganisms in a natural community. *Nature*, **344**, 33–44.

Wang,G.C.T. and Wang,Y. (1996) The frequency of chimeric molecules as a consequence of PCR co-amplification of 16S rRNA genes from different bacterial species. *Microbiology*, **142**, 1107–1114.